Claims

What is claimed is:

1. A method of classifying substructures of at least one unmarked string using at least one training data set, comprising the steps of:

constructing a model of class labels and substructures within strings of the at least one training data set, wherein the training data set has markers identifying labeled substructures;

inserting markers in the at least one unmarked string in accordance with the model, wherein the markers identify substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set; and

predicting class labels of substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set in accordance with the model.

2. The method of claim 1, wherein each training data set comprises:

a set of strings, wherein each string comprises at least one marker identifying a beginning of a substructure and at least one marker identifying an end of a substructure; and

a set of class labels for association with the substructures.

3. The method of claim 2, wherein each substructure within each training data set comprises:

a begin marker indicating a beginning of the substructure;

an end marker indicating an end of the substructure; and

a class label indicating a class of the substructure.

4. The method of claim 1, wherein the at least one unmarked string comprises a first inner substructure nested within a second outer substructure.

5. The method of claim 1, wherein the at least one unmarked string comprises sequentially occurring substructures.

6. The method of claim 1, wherein the at least one unmarked string comprises substructures having different class labels.

7. The method of claim 1, wherein the step of constructing a model comprises modeling a relationship of substructures in strings of the at least one training data set using a stochastic model.

8. The method of claim 7, further comprising the step of generating a symbol having a probability of performing a transition from one of a number of states to another in the stochastic model.

9. The method of claim 3, wherein a model for a simple structural embedding comprises three sets of states: (a) a set of states for a portion of the string before the begin marker; (b) a set of states for a portion of the string within the begin marker and end marker; and (c) a set of states for a portion of the string after the end marker.

10. The method of claim 4, wherein a model for the first inner substructure nested within the second outer substructure has five sets of states: (a) a set of states for a portion of the string before both the inner substructure and the outer substructure; (b) a set of states for a portion of the string within the outer substructure, but before the inner substructure; (c) a set of states for a portion of the string within both the inner

substructure and the outer substructure; (d) a set of states for a portion of the string within the outer substructure, but after the inner substructure; and (e) a set of states for a portion of the string after both the inner substructure and the outer substructure.

11. The method of claim 5, wherein the model for sequentially occurring substructures within a string has different sets of states for each sequential substructure.

12. The method of claim 6, wherein the model for the at least one unmarked string with substructures having different class labels has a different set of states for each class label.

13. The method of claim 7, wherein each string corresponds to a path in the stochastic model.

14. The method of claim 13, wherein a probability of the path corresponds to a level of correlation of the model to the string.

15. The method of claim 8, further comprising the step of determining the number of states in the model using the at least one training data set.

16. The method of claim 15, wherein the number of states in the model is a function of a length of the strings in the model.

17. The method of claim 8, further comprising the step of determining the probability of transition from one state to another using the at least one training data set.

18. The method of claim 17, wherein the probability of transition from one state to another is a function of a percentage of transition between the states using the at least one training data set.

19. The method of claim 8, further comprising the step of determining the probability of generation of the symbols at a given state using the at least one training data set.

20. The method of claim 19, wherein the step of determining the probability of generation of the symbols at a given state comprises using a probability of generation of each symbol at a given state in the at least one training data set.

21. The method of claim 1, wherein the step of inserting markers in the at least one unmarked string comprises inserting markers in such a way that a corresponding path in a stochastic model has a maximum probability.

22. Apparatus for classifying substructures of at least one unmarked string, using at least one training data set, comprising:

a memory; and

at least one processor, coupled to the memory, operative to: (i) construct a model of class labels and substructures within strings of the at least one training data set, wherein the training data set has markers identifying labeled substructures; (ii) insert markers in the at least one unmarked string in accordance with the model, wherein the markers identify substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set; and (iii) predict class labels of substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set in accordance with the model.

23. The apparatus of claim 22, wherein each training data set comprises:

a set of strings, wherein each string comprises at least one marker identifying a beginning of a substructure and at least one marker identifying an end of a substructure; and

a set of class labels for association with the substructures.

24. The apparatus of claim 23, wherein each substructure within each training data set comprises:

a begin marker indicating a beginning of the substructure;

an end marker indicating an end of the substructure; and

a class label indicating a class of the substructure.

25. The apparatus of claim 22, wherein the at least one unmarked string comprises a first inner substructure nested within a second outer substructure.

26. The apparatus of claim 22, wherein the at least one unmarked string comprises sequentially occurring substructures.

27. The apparatus of claim 22, wherein the at least one unmarked string comprises substructures having different class labels.

28. The apparatus of claim 22, wherein the operation of constructing a model comprises modeling a relationship of substructures in strings of the at least one training data set using a stochastic model.

29. The apparatus of claim 28, wherein the at least one processor is further operative to generate a symbol having a probability of performing a transition from one of a number of states to another in the stochastic model.

30. The apparatus of claim 24, wherein a model for a simple structural embedding comprises three sets of states: (a) a set of states for a portion of the string before the begin marker; (b) a set of states for a portion of the string within the begin marker and end marker; and (c) a set of states for a portion of the string after the end marker.

31. The apparatus of claim 25, wherein a model for the first inner substructure nested within the second outer substructure has five sets of states: (a) a set of states for a portion of the string before both the inner substructure and the outer substructure; (b) a set of states for a portion of the string within the outer substructure, but before the inner substructure; (c) a set of states for a portion of the string within both the inner substructure and the outer substructure; (d) a set of states for a portion of the string within the outer substructure, but after the inner substructure; and (e) a set of states for a portion of the string after both the inner substructure and the outer substructure.

32. The apparatus of claim 26, wherein the model for sequentially occurring substructures within a string has different sets of states for each sequential substructure.

33. The apparatus of claim 27, wherein the model for the at least one unmarked string with substructures having different class labels has a different set of states for each class label.

34. The apparatus of claim 28, wherein each string corresponds to a path in the stochastic model.

35. The apparatus of claim 34, wherein a probability of the path corresponds to a level of correlation of the model to the string.

36. The apparatus of claim 29, wherein the at least one processor is further operative to determine the number of states in the model using the at least one training data set.

37. The apparatus of claim 36, wherein the number of states in the model is a function of a length of the strings in the model.

38. The apparatus of claim 29, wherein the at least one processor is further operative to determine the probability of transition from one state to another using the at least one training data set.

39. The apparatus of claim 38, wherein the probability of transition from one state to another is a function of a percentage of transition between the states using the at least one training data set.

40. The apparatus of claim 29, wherein the at least one processor is further operative to determine the probability of generation of the symbols at a given state using the at least one training data set.

41. The apparatus of claim 40, wherein the operation of determining the probability of generation of the symbols at a given state comprises using a probability of generation of each symbol at a given state in the at least one training data set.

42. The apparatus of claim 22, wherein the operation of inserting markers in the at least one unmarked string comprises inserting markers in such a way that a corresponding path in a stochastic model has a maximum probability.

43. An article of manufacture for classifying substructures of at least one unmarked string, using at least one training data set, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

constructing a model of class labels and substructures within strings of the at least one training data set, wherein the training data set has markers identifying labeled substructures;

inserting markers in the at least one unmarked string in accordance with the model, wherein the markers identify substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set; and

predicting class labels of substructures in the at least one unmarked string similar to substructures within strings of the at least one training data set in accordance with the model.